Additional File 3: Training results obtained using various combinations of 'cumulative sequence count' (20K, 30K, 40K, 50K, 60K, 70K, 80K) and 'overlap percentage' (20%, 30%, 40%, 50%, 60%, 70%, 80%). The four tables A-D show results obtained with Sanger, 454-400, 454-250 and 454-100 training data sets respectively.

Note:
Cumulative sequence count: The total number of sequences in a set of clusters identified as having the least distance

Overlap percentage: The percentage of overlap between the set of clusters (identified as compositionally closest to a
## (A) Sanger training data set

Total Number of sequences: 550,000 (10,000 reads each from 55 training organisms)
Sequences harboring portions of the 16 rDNA gene: 1411 (0.26%)

<table>
<thead>
<tr>
<th>Overlap percentage*</th>
<th>A and B values**</th>
<th>Value of cumulative sequence count*** used for identifying a set of clusters closest to the query sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>20K</td>
</tr>
<tr>
<td>20%</td>
<td>A</td>
<td>31.5</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>89.9</td>
</tr>
<tr>
<td>30%</td>
<td>A</td>
<td>24.5</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>87.5</td>
</tr>
<tr>
<td>40%</td>
<td>A</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>83.2</td>
</tr>
<tr>
<td>50%</td>
<td>A</td>
<td>7.2</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>74.6</td>
</tr>
<tr>
<td>60%</td>
<td>A</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>69.9</td>
</tr>
<tr>
<td>70%</td>
<td>A</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>51.4</td>
</tr>
<tr>
<td>80%</td>
<td>A</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>48.7</td>
</tr>
</tbody>
</table>

* Overlap percentage: The percentage of overlap between the set of clusters (identified as compositionally closest to a given query sequence) and the set of pre-tagged 'probable 16S rDNA clusters'

** A represents the percentage of training data set sequences reported by i-rDNA as 'probable' 16S rDNA fragments

** B represents the percentage of 'true' 16S rDNA fragments within the subset of sequences reported by i-rDNA as probable 16S rDNA fragments.

***Cumulative sequence count: The total number of sequences in a set of clusters identified as having the least distance to the vector corresponding to a query sequence.
(B) 454-400 training data set
Total Number of sequences: 550,000 (10,000 reads each from 55 training organisms)
Sequences harboring portions of the 16 rDNA gene: 765 (0.14%)

<table>
<thead>
<tr>
<th>Overlap percentage*</th>
<th>A and B values**</th>
<th>Value of 'cumulative sequence count' used for identifying a set of clusters closest to the query sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>20K</td>
</tr>
<tr>
<td>20%</td>
<td>A</td>
<td>33.2</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>92.4</td>
</tr>
<tr>
<td>30%</td>
<td>A</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>90.5</td>
</tr>
<tr>
<td>40%</td>
<td>A</td>
<td>17.9</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>89.8</td>
</tr>
<tr>
<td>50%</td>
<td>A</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>86.4</td>
</tr>
<tr>
<td>60%</td>
<td>A</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>61.3</td>
</tr>
<tr>
<td>70%</td>
<td>A</td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>54.1</td>
</tr>
<tr>
<td>80%</td>
<td>A</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>45.4</td>
</tr>
</tbody>
</table>

* Overlap percentage: The percentage of overlap between the set of clusters (identified as compositionally closest to a given query sequence) and the set of pre-tagged ‘probable 16S rDNA clusters’

** A represents the percentage of training data set sequences reported by i-rDNA as ‘probable’ 16S rDNA fragments

** B represents the percentage of ‘true’ 16S rDNA fragments within the subset of sequences reported by i-rDNA as probable 16S rDNA fragments.

***Cumulative sequence count: The total number of sequences in a set of clusters identified as having the least distance to the vector corresponding to a query sequence.
### (C) 454-250 training data set

**Total Number of sequences:** 550,000 (10,000 reads each from 55 training organisms)

Sequences harboring portions of the 16 rDNA gene: 720 (0.13%)

<table>
<thead>
<tr>
<th>Overlap percentage*</th>
<th>A and B values**</th>
<th>Value of ‘cumulative sequence count’ used for identifying a set of clusters closest to the query sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>20K</td>
</tr>
<tr>
<td>20%</td>
<td>A</td>
<td>34.2</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>87.1</td>
</tr>
<tr>
<td>30%</td>
<td>A</td>
<td>19.5</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>84.4</td>
</tr>
<tr>
<td>40%</td>
<td>A</td>
<td>14.5</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>81.2</td>
</tr>
<tr>
<td>50%</td>
<td>A</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>80.5</td>
</tr>
<tr>
<td>60%</td>
<td>A</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>74.5</td>
</tr>
<tr>
<td>70%</td>
<td>A</td>
<td>4.25</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>43.7</td>
</tr>
<tr>
<td>80%</td>
<td>A</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>33.7</td>
</tr>
</tbody>
</table>

* Overlap percentage: The percentage of overlap between the set of clusters (identified as compositionally closest to a given query sequence) and the set of pre-tagged ‘probable 16S rDNA clusters’

** A represents the percentage of training data set sequences reported by i-rDNA as ‘probable’ 16S rDNA fragments

** B represents the percentage of ‘true’ 16S rDNA fragments within the subset of sequences reported by i-rDNA as probable 16S rDNA fragments.

***Cumulative sequence count: The total number of sequences in a set of clusters identified as having the least distance to the vector corresponding to a query sequence.
## (D) 454-100 training data set

Total Number of sequences: 550,000 (10,000 reads each from 55 training organisms)
Sequences harboring portions of the 16 rDNA gene: 461 (0.08%)

<table>
<thead>
<tr>
<th>Overlap percentage*</th>
<th>A and B values**</th>
<th>Value of 'cumulative sequence count' used for identifying a set of clusters closest to the query sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>20%</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>30%</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>40%</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>60%</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>70%</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>80%</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
</tr>
</tbody>
</table>

* Overlap percentage: The percentage of overlap between the set of clusters (identified as compositionally closest to a given query sequence) and the set of pre-tagged 'probable 16S rDNA clusters'

** A represents the percentage of training data set sequences reported by i-rDNA as 'probable' 16S rDNA fragments

** B represents the percentage of 'true' 16S rDNA fragments within the subset of sequences reported by i-rDNA as probable 16S rDNA fragments.

***Cumulative sequence count: The total number of sequences in a set of clusters identified as having the least distance to the vector corresponding to a query sequence.